Supporting information about constructing student outcome and experience indicators for use in OfS regulation

Description of statistical methods

Enquiries to providermetrics@officeforstudents.org.uk

Publication date 20 January 2022

Contents

Introduction

Purpose

1. The Office for Students (OfS) has issued a consultation about the construction of student outcome and experience measures to be used in our regulation of student outcomes and the Teaching Excellence Framework (TEF).[1] This document has been published as supporting information alongside the consultation, to aid higher education providers and other stakeholders in understanding the range of statistical methods we have proposed to use in the presentation and contextualisation of the indicators. We anticipate that some readers of the consultation proposals, particularly those with in-depth knowledge of statistical methods, will find the information in this supporting document useful for exploring the practical effects of implementing our proposals.

2. The statistical methods described in this document are directly aligned to our consultation proposals and remain subject to change upon conclusion of the consultation exercise. They have been formulated on the same basis as described in our consultation on the construction of student outcome and experience indicators.[2] We expect to publish similar information to that found in this document alongside publication of the consultation outcomes later in 2022, and when we implement our final approach to constructing student outcome and experience data indicators.

3. The statistical methods described in this document have only been proposed for use in respect of indicators constructed to inform our regulation of student outcomes and the TEF. We do not at this stage intend to make use of these approaches in other OfS publications of student outcome and experience indicators. For example, the statistical methods that have been employed in the access and participation data dashboard have been described in technical documentation published on the OfS website and at this stage remain unchanged.[3]

Guidance for using this document

4. This is one of a series of supporting technical documents that provide details of the definitions and methods that the OfS has proposed to use in constructing student outcome and experience data indicators. Readers may want to consider this document alongside the following documents and resources in particular:

[1] See www.officeforstudents.org.uk/publications/student-outcomes-and-teaching-excellence-consultations/student-outcomes-data-indicators/.

[2] See proposals 10 and 11 of the consultation on constructing student outcome and experience indicators for use in OfS regulation, at www.officeforstudents.org.uk/publications/student-outcomes-and-teaching-excellence-consultations/student-outcomes-data-indicators/.

[3] Regulatory indicators, methodology and rebuild instructions, available at www.officeforstudents.org.uk/data-and-analysis/institutional-performance-measures/technical-documentation/.

Consultations[4]

- Consultation on regulating student outcomes, in particular proposals 5 and 6.
- Consultation on the TEF, in particular proposals 9 and 11.
- Consultation on constructing student outcome and experience indicators for use in OfS regulation, in particular proposals 10 and 11.

Supporting information about constructing student outcome and experience indicators for use in OfS regulation[5]

- Student outcomes data dashboard
- TEF data dashboard
- Dashboard user guide
- Review of the selection and grouping of benchmarking factors.[6]

5. This document is split into two broad topics, with further technical detail on each topic included in the annexes:

   a. The presentation of student outcome and experience data indicators and communicating statistical uncertainty.

   b. The approach to contextualise the student outcome and experience data indicators through benchmarking.

Enquires and feedback

6. Enquiries about the methods described in this document should be sent to providermetrics@officeforstudents.org.uk.

[4] All available at www.officeforstudents.org.uk/publications/student-outcomes-and-teaching-excellence-consultations/.

[5] All available at www.officeforstudents.org.uk/data-and-analysis/student-outcomes-and-experiences-data-dashboards/.

[6] Available at www.officeforstudents.org.uk/publications/student-outcomes-and-teaching-excellence-consultations/outcome-and-experience-data/.

Presentation of the indicators and statistical uncertainty

What is statistical uncertainty?

7. When calculating student outcome and experience measures as data indicators, each indicator that the OfS calculates is a factual representation of the outcomes or experiences of students observed at a particular provider at a particular point in time. If one is interested only in the actual population of students present at a particular provider at a particular time, then it would be appropriate to rely solely on this value.

8. The group of students which actually did attend are just one realisation of many other populations of students who could have attended that provider or may do so in the future. The observed population is – in various respects – a random realisation of those other populations. If that realisation had been different, for example if the observed population at the provider had included a few more ‘morning people’ and a few fewer ‘night owls’, would attendance at lectures have had a different influence over continuation or completion outcomes? If it happened to be raining on the day that students chose to complete a survey, how differently would student experiences be reported in comparison with the responses that would have been made if it happened to be sunny instead? This randomness could give rise to a slight difference in the observed population that could give rise to slightly different indicator values being calculated, even though the underlying performance of the provider and their course delivery remained the same. This potential for random variation in the indicator values we calculate is known as statistical uncertainty.

Why is uncertainty important?

9. Within the OfS’s regulatory uses of student outcome and experience indicators, we want to think about indicator values as representing something about a whole population of students who could have attended that provider, or may do so in the future. This whole population is known as a superpopulation. The group which actually did attend are just one possible set of students from this superpopulation, and the value calculated from data about the set of students which actually did attend is used as an estimate for what we would expect in the superpopulation.

10. As described in paragraph 9 above, in theory, we could have looked at data from a different set of students from the superpopulation, and this could have given a slightly different answer: any of those answers could be used as an estimate for the value in the superpopulation. Of course, in practice, we are not able to look at data from a different set of students from the superpopulation. Students who could have attended the provider in question but did not do so, and students who may attend the provider in future, cannot be known to us and do not exist in the student data available to us. The term statistical inference is used to describe the process of using data about one thing we know about, to infer what might happen in a similar situation where we don’t have full data. In this case, we are using data about one set of students to infer what we would expect in the superpopulation.

11. As such, there will always be a question as to how exact any calculated indicator value is as an estimate for the superpopulation. This question of exactness (or of statistical uncertainty), and the notion that indicator values may not be precise measures of the underlying performance

that they aim to represent, is important because we have proposed to use these indicator values within the OfS's regulatory uses of student outcome and experience indicators to help us to understand the underlying true performance of a provider in respect of the outcomes and experiences it delivers for students. We acknowledge that it is not possible to say exactly what a provider's underlying performance looks like for the superpopulation.

12. Any judgement of performance is a judgement about the superpopulation so should be aware of the potential extent of this statistical uncertainty. Identifying meaningful and effective ways to quantify and communicate the potential extent of statistical uncertainty is therefore essential in the context of our proposed uses.

Statistical uncertainty, not measurement error

13. Statistical uncertainty should not be confused for measurement error (sometimes known as observational error). Measurement error occurs when there are inaccuracies either in the underlying data on which we are performing our calculations (for example, a student is erroneously reported as studying full-time rather than part-time), or within the calculations that we are performing (for example, a formula that should include a 'greater than or equals to' condition mistakenly includes a 'strictly greater than' condition instead).

14. While neither example of measurement error can be entirely ruled out, we aim to identify and rectify any such errors through our sharing of the data and methods used with providers and other stakeholders. We welcome feedback on any methodological oversights we may have made. However, we are confident that the indicators we have calculated are an accurate factual representation of student outcomes and experiences as they have been reported to us through the student data returns that inform those indicators.

15. Statistical uncertainty is unavoidable in the calculation of any statistic that is unable to identify and refer to the superpopulation: it cannot be rectified through adjustments to the underlying data or the calculations we are performing. It therefore requires explicit consideration in our presentation and communication of student outcome and experience data indicators, and in our assessments thereof.

General approach to presenting uncertainty

16. As a producer of official statistics, the OfS is committed to effectively communicating its statistics, to allow users to assess and have confidence in the value of the statistics and avoid misinterpretation of them. This, together with the use of these indicators to inform our regulation of student outcomes and in the TEF, as well as in regulation of access and participation, means that we take the view that it is essential to identify meaningful and effective ways to provide an awareness of the potential extent of statistical uncertainty.

17. When presenting student outcome and experience indicators to inform our regulation of student outcomes and the TEF, we have chosen to use 'shaded bars' to represent the statistical uncertainty associated with observed values. There are two observed values that have been proposed for use in these assessments, and we will show a shaded bar in respect of each case:

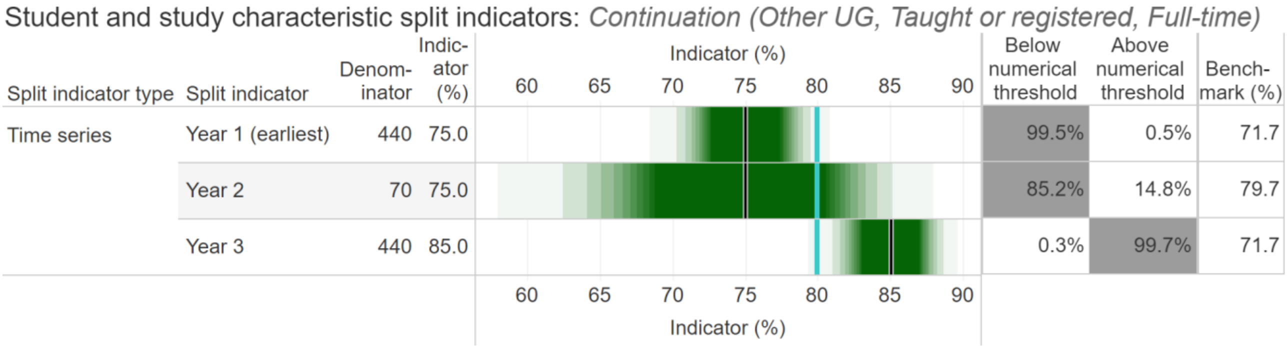
    a. The observed value of the **indicator** as a point estimate, reporting the proportion of students that we observe to have achieved a certain outcome or reported a certain experience. Throughout the remainder of this document we refer to this as a measure of the provider's **absolute** performance.

b. The observed value of the **difference** between the indicator and its associated benchmark, as a point estimate. Throughout the remainder of this document we refer to this as a measure of the provider's **relative** performance.

18. The shaded bars that we are showing are illustrated below in Figures 1 and 2. They aim to represent the continuous spread (or distribution) of statistical uncertainty around the point estimates that we have calculated. The shading of the bars indicates the changing likelihood that underlying provider performance takes different values, with the darkest shading representing the range in which there is the greatest likelihood that true underlying provider performance might lie. Much like the bell curve of a normal distribution, as the shading lightens in both directions it represents a lower likelihood that true underlying performance falls at that point. Wider shaded bars mean that we become less confident in the observed point estimate.

19. The two bars are differentiated by colour, to represent the different interpretations of performance. The spread of statistical uncertainty associated with the absolute performance is represented in a green shaded bar, whereas that associated with the relative performance is represented in a blue shaded bar.

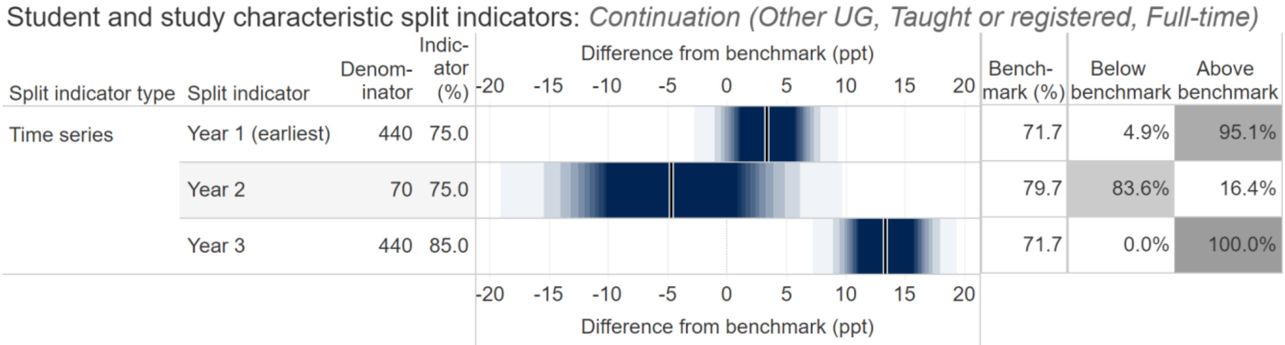
Figure 1: Example of green shaded bars, showing spread of statistical uncertainty around indicator values



Student and study characteristic split indicators: *Continuation (Other UG, Taught or registered, Full-time)*

Split indicator type	Split indicator	Denom-inator	Indic-ator (%)	Indicator (%)	Below numerical threshold	Above numerical threshold	Bench-mark (%)
Time series	Year 1 (earliest)	440	75.0		99.5%	0.5%	71.7
	Year 2	70	75.0		85.2%	14.8%	79.7
	Year 3	440	85.0		0.3%	99.7%	71.7

Figure 2: Example of blue shaded bars, showing spread of statistical uncertainty around difference between indicator and benchmark



Student and study characteristic split indicators: *Continuation (Other UG, Taught or registered, Full-time)*

Split indicator type	Split indicator	Denom-inator	Indic-ator (%)	Difference from benchmark (ppt)	Bench-mark (%)	Below benchmark	Above benchmark
Time series	Year 1 (earliest)	440	75.0		71.7	4.9%	95.1%
	Year 2	70	75.0		79.7	83.6%	16.4%
	Year 3	440	85.0		71.7	0.0%	100.0%

20. The presentation of the shaded bars is intentionally similar and, in broad terms, each can be thought of as representing a series of discrete confidence intervals around the point estimate we have observed, where each confidence interval in the series corresponds to a different confidence (or significance) level.

What is a confidence interval?

21. One way in which statistics can help to describe the level of statistical uncertainty associated with a point estimate calculated from the observed data is to supply a range of reasonable values for a provider's underlying performance in the superpopulation. This range of reasonable values is called a confidence interval.

22. A confidence interval has an associated confidence level, which represents the likelihood that the true value of underlying performance would be contained within that proportion of confidence intervals computed in relation to the superpopulation. In other words, 95 per cent of confidence intervals computed at the 95 per cent confidence level would contain the true value of performance in the superpopulation, 90 per cent of confidence intervals computed at the 90 per cent confidence level would contain the true value, and likewise for other confidence levels.

23. This means that the width of the confidence interval is influenced by the desired confidence level. It is also influenced by the number of students informing the calculation of the point estimate from the observed data (otherwise known as the sample size), where, as the number of students increases, the width of the confidence interval tends to decrease. The variability in the sample – the consistency of the observed student outcomes or experiences – can also influence the width of the interval, with more variable samples generating wider confidence intervals. Wider confidence intervals mean that we become less confident in the observed point estimate.

24. In common parlance, for most proportions, a confidence interval will often be expressed as plus or minus a similar amount, such as '50 per cent plus or minus 3 percentage points'. In extreme cases where the observed point estimate is very large or very small (for example, close to 100 per cent or to zero per cent), it is theoretically possible for a calculated confidence interval to extend above 100 per cent or below zero per cent. The OfS does not report limits of confidence intervals that are above 100 per cent or below zero per cent. In such cases, when it is clearly impossible for the proportion to actually fall below 0 per cent or above 100 per cent, confidence intervals can appear truncated at one end and not be symmetrical.

How are we using confidence intervals in the shaded bars?

25. In designing the shaded bars, we have sought to avoid selecting a single confidence interval significance level. To do so would create a 'cliff edge' at a single significance level pre-determined by the OfS for our specific use, which would facilitate a binary interpretation of performance as definitively above or below a given threshold by most users. Instead, we illustrate the distribution of statistical uncertainty up to a maximum of a 99.7 per cent confidence interval and have proposed that our own assessments of a provider's performance will establish the statistical confidence we have in relation to its performance by considering the uncertainty distribution relative to our proposed numerical thresholds. We also anticipate that other users of the data will be empowered to better understand the confidence in which they can hold their own judgements of student outcomes and experiences.

26. Our construction of the shaded bars requires a set of assumptions to be made about the statistical distributions from which the statistics are drawn. These assumptions, and their resulting influence over the methods we have selected, vary in respect of our consideration of absolute and relative performance and are explained in greater detail in Annexes A and B respectively.

27. The shaded bars are constructed around the point estimate by calculating a set of confidence intervals, starting with the 75 per cent confidence interval with further intervals calculated at 2.5 percentage point increments up to a maximum of a 99.7 per cent confidence interval. The bar is shaded between each of these intervals to represent the shape of the underlying distribution, with the darkest shading representing the range in which has the highest likelihood that true underlying provider performance might lie. We illustrate the distribution up to a maximum of a 99.7 per cent confidence interval. This approach means that we maximise the chance that the shaded bars encapsulate the true underlying performance, and that users are empowered to better understand the confidence in which they can hold their own judgements of student outcomes and experiences by making their own choice of confidence intervals.

28. Our regulation of student outcomes and the TEF will make use of these shaded bars by establishing the confidence with which we can say that the underlying performance provider is above or below a given numerical value. To facilitate consistent interpretations of this confidence, we have summarised the proportion of the distribution represented by the shaded bar that falls above or below those values. These summary figures are reported in a supplementary table alongside the shaded bars, with the intention that the two are used together to inform an accurate and consistent interpretation of statistical confidence, related to the numerical values that the OfS have proposed to make use of. These summary figures are highlighted where they report that at least 75 per cent of the distribution represented by the shaded bar falls above or below those values. Users can of course use the shaded bars to make other interpretations of the performance we are representing. These are illustrated in Figures 1 and 2 above. The calculations underpinning these summary figures are provided for each type of shaded bar in Annexes A and B respectively.

Multiple comparison adjustments

29. When multiple statistics are calculated on a given topic, it is often expected that users will wish to make comparisons between those statistics. To the extent that those statistics include information about statistical uncertainty, that uncertainty can be underestimated depending on the nature of the multiple comparisons that are being made. For instance, in the case of 95 per cent confidence intervals, the likelihood that the computed confidence interval includes the true value of underlying performance may be substantially lower than the intended 95 per cent if multiple comparisons are being made. For example, if you were making a comparison of the statistical significance at the 95 per cent confidence interval across the performance of 20 subjects at a provider without a multiple comparison adjustment, on average one subject (5 per cent of 20 subjects) appears to be statistically significant but is not in fact significant. To overcome this, adjustments can be made to the calculations to control the error or false discovery rates (such as the Bonferroni correction).

30. We have proposed not to make any such adjustments for multiple comparisons within our construction of student outcome and experience indicators to inform our regulation of student outcomes and the TEF:

a. Our use of the shaded bars in the presentation of the data aims to portray the distribution of statistical uncertainty and does not rely on a single confidence interval or significance test. We consider that the presentation of uncertainty up to the 99.7 per cent confidence interval is already broadly sufficient to encapsulate the true underlying performance.

b. The number of comparisons that users might make within and across the full set of available data points is very large and unpredictable, and likely to vary by user. This might lead us to a substantial adjustment for multiple comparisons, which we do not consider would be proportionate given a. above.

31. While we have proposed not to adjust for multiple comparisons, we do ask users who wish to make multiple comparisons to consider adjusting to a higher level of confidence when making their judgements because of the higher risk of false discovery when using lower levels of statistical confidence. Users should be more conservative in their interpretation of statistical uncertainty the more comparisons they are making.

Benchmarking

32. The OfS has proposed to use benchmarking to inform our regulation of student outcomes and the TEF, to help interpret a provider's actual performance relative to that in the sector overall once we have taken into account the mix of students at the provider or the provision being offered. Each indicator that the OfS calculates represents the outcomes that we have observed for the students at a particular provider at a particular point in time. The calculation of a benchmark gives us a counterfactual for the observed outcomes, which we intend can be used in two ways:

    a. to understand a provider's performance in relation to the higher education sector as a whole

    b. to assess similarities between individual providers.

33. In making these comparisons, we take account of factors which describe the profile of students and provision delivered by higher education providers and which are correlated with the outcomes we are measuring. The benchmarking methodology used by the OfS involves consideration of unique combinations of the student and course characteristics that we have selected to act as benchmarking factors: we refer to these unique combinations as benchmarking groups.

34. The methodology allows us to ask the question: "What would the observed student outcome have been at this provider if its distribution of students across benchmarking factor groups had been what it was, but its outcomes across those same benchmarking groups were replaced by the sector-overall rates?".

35. When there are known differences between the outcomes and experiences of some groups of students or providers, observed average values for the whole of the higher education sector are not necessarily helpful when forming this expectation. Instead, we calculate the benchmark as a weighted sector average reflecting the number of students in that group at the provider. As such, benchmarks give information about the values that the sector overall might have achieved for the indicator if the characteristics included in the benchmarking factors are the only ones that are important. Where differences exist between an indicator and its corresponding benchmark, these may be due to the provider's performance, or they may be due to some other characteristic which is not included in the weighting.

General approach to benchmarking

36. To create benchmarks, we calculate the observed rates for the higher education sector as a whole for each benchmarking group. The benchmark for each provider is then calculated by taking a weighted average of the overall sector outcomes for each benchmarking group, taking account of the particular mix of students across those groups at the provider in question. A worked example is provided in Annex D.

37. The benchmarking methodology used by the OfS means that a provider is not being compared with a pre-set group of providers, but rather the outcomes for a provider's students are compared with the outcomes of similar students across the entirety of the higher education

sector. For the purpose of calculating these benchmarks, the higher education sector within which we are making comparisons of the outcomes for similar students is made up of:

a. **For OfS registered providers:** all English higher education providers registered with the OfS at the time that we produce the indicators.

b. **For providers in the devolved administrations:** all English higher education providers registered with the OfS and all those providers who are funded or regulated by one of the devolved administration organisations at the time that we produce the indicators.

The benchmarking factors we use

38. The basis on which we proposed to select, define, and apply the factors used in benchmarking student outcome and experience indicators is key to the integrity and robustness of the benchmark values calculated and assessed. Our selection and application of benchmarking factors is underpinned by a set of guiding principles.[7] The benchmarking factors used for each measure, with variations applied for providers in the devolved administrations, are described in our consultation on constructing student outcome and experience indicators for use in OfS regulation.[8]

Risks of self-benchmarking

39. When constructing the benchmark for an individual provider, the students at that provider contribute to the sector averages we calculate. We recognise that where the characteristics of students at the provider in question do not frequently occur among student populations in the wider sector, these sector averages may be heavily influenced by that provider. This is referred to as the risk of 'self-benchmarking'. In such a scenario, the provider's own students would be making a substantial contribution to the calculation of its benchmark, making the calculation less robust and the resulting benchmark value less meaningful. The benchmark value will become more similar to the indicator value as the provider's contribution increases. This is because there is little other sector data that can provide the information necessary to make the benchmark a reliable estimate of the values that might have been expected for the provider.

40. The risk of self-benchmarking becomes more acute when benchmarking groups are defined at such a detailed level that only very small numbers of students possess each unique combination of the student and course characteristics that we have selected to act as benchmarking factors. When many benchmarking groups are populated by only one or two students, the sector averages calculated for those groups will tend to a small range of values. If the sector average is calculated in reference to a single student, it can only result in an 'average' of either 0 per cent or 100 per cent. If it refers to only two students, the average can only be 0 per cent, 50 per cent or 100 per cent. Sector averages that include large numbers of 0 per cent and 100 per cent values can lead to an ineffectual weighting which will skew the

[7] See Annex D of the consultation on constructing student outcome and experience indicators for use in OfS regulation, at www.officeforstudents.org.uk/publications/student-outcomes-and-teaching-excellence-consultations/student-outcomes-data-indicators/.

[8] See proposal 10 of the consultation on constructing student outcome and experience indicators for use in OfS regulation, at www.officeforstudents.org.uk/publications/student-outcomes-and-teaching-excellence-consultations/student-outcomes-data-indicators/.

resulting benchmark and increase the standard errors of the calculated difference between indicator and benchmark values.[9]

41. Our proposed selection of benchmarking factors has sought to minimise the occasions on which we might encounter self-benchmarking, by selecting and grouping factors in such a way as to ensure as far as possible that reasonable numbers of students from multiple providers are contributing to each sector average that we calculate. We are aware that the diversity of the higher education sector means that we cannot mitigate this risk entirely and our proposed benchmarking factors tolerate a risk of self-benchmarking on a small scale. To facilitate an understanding of where this situation may occur, we propose to include information about the provider's own contribution to that benchmark within the datasets we construct. This will also support users of the information in the public domain to understand and respond to the risk that the benchmark is of limited use. Our calculation of a provider's own contribution to its benchmark is explained further in Annex C.

Benchmarking split indicators

42. The approach to benchmarking split indicators mostly follows the general approach described in paragraphs 36 and 37. However, instead of creating a benchmark for the provider using data from every provider in the sector, we repeat that process per split indicator and subset the provider and the sector to the split indicator in question. This approach is equivalent to including the definition of the split as a benchmarking factor for each split. For example, to benchmark the 'Male' split indicator we subset the provider and the sector to only male students, so that we can compare the student outcomes for male students at the provider to a benchmark created from male students across the sector. We then separately benchmark the 'Female' split indicator by sub-setting the provider and the sector to only female students. This approach can lead to cases such as where a provider's relative performance could appear below benchmark for the all years aggregate split, but appear above benchmark for every other split.

Benchmarking suppression

43. Some of the factors proposed as benchmarking factors are known to include attributes identifying the characteristic or information as unknown, not required or not applicable. This occurs where student data has not been returned for the OfS to be able to classify students appropriately, whether because this information was not shared with a provider, so it has been unable to include it in its HESA or Individualised Learner Record (ILR) data submissions, or because those data returns do not currently require the collection of that information.

44. A large number of students being reported with unknown attributes reported for a benchmarking factor can impact on the reliability of the benchmarking calculations. Our benchmarking method is effective in taking account of the mix of a provider's students and provision when the grouping of attributes within benchmarking factors forms coherent groups which share a consistency of student backgrounds, outcomes, or behaviours with respect to the indicator to which they refer. By virtue of the attribute being reported as unknown, we cannot know the extent to which students reported in this way actually do form coherent,

[9] The standard errors of a statistic represent the amount by which one would expect that statistic to change, based solely on random sampling.

homogeneous groups, nor the extent to which weighting the sector average for the size of this group becomes akin to comparing apples and pears. We therefore take the view that a large number of students being reported with unknown attributes dilutes the effect of that characteristic on the efficiency of the calculated benchmark.

45. Our proposed definitions of the benchmarking factors have sought to mitigate this risk through our adoption of the proposed guiding principles for the selection and application of benchmarking factors. However, an individual provider’s benchmark will still be impacted by this risk if significant numbers of unknown attributes are returned for those factors in their student data.

46. We have therefore proposed to suppress a benchmark value where a provider’s student data reports at least 50 per cent of the students with unknown information for one or more of the factors used for that benchmark calculation. We consider that there is insufficient data to form reliable benchmarks when a majority of students at the provider have unknown information for at least one of the benchmarking factors. For example, where entry qualifications are proposed as a benchmarking factor, the benchmark value (and the calculated difference between the indicator and the benchmark) is suppressed if at least 50 per cent of the provider’s students have unknown entry qualifications.

Adjustments to the general approach to benchmarking

For the ‘taught or registered (TorR)’ population

47. The indicators are constructed for several different views of a provider’s student population. The general approach to benchmarking can be applied to each of the registered and taught populations. We have proposed to use information about the population of students taught at a given provider in our regulation of student outcomes and will therefore apply the general benchmarking approach to the construction of those indicators.

48. We have also proposed that in our regulation of student outcomes and the TEF, we will look at the population of students who are either registered or taught at the provider in question. This requires an adjustment to the approach used to construct benchmarks in this view. This is because students can be associated with more than one provider in the indicators. However, the benchmarking methodology assumes that students per provider per unique combination of benchmarking factors are independent from another combination. The benchmarking calculations for the taught or registered views are therefore adjusted as described in Annex C, to accommodate the potential for a student to contribute to the indicators and benchmarks of multiple providers.

For the compound completion measure

49. Whilst the general approach to benchmarking applies to all measures of student outcomes and experiences, it relies on the measure being individual-based, which the compound completion measure is not. To construct the benchmark for the compound completion measure we calculate the observed withdrawal rates for the higher education sector for each benchmarking group for each of the six entry cohorts that are used to construct the cohort-based measure. In doing so we can treat this like an individual-based measure. This method is explained in greater detail in Annex C.

Annex A: Presenting uncertainty about absolute performance

1. In presenting student outcomes and student experience indicators to inform our regulation of student outcomes and the TEF, a provider's absolute performance is represented in green shaded bars. They represent the statistical uncertainty associated with the observed value of the **indicator** as a point estimate. This annex provides a fuller technical description of the statistical methods used to compute the confidence intervals which contribute to the construction of the green shaded bars. It is aimed at readers with an in-depth knowledge of advanced statistical methods and assumes a familiarity with statistical formulae and notation.

2. Typically for this type of observed outcome, you would create a binomial proportion confidence interval, where the probability of success and the number of trials is given by the observed indicator value and the number of students informing the indicator respectively (the denominator).

General approach

3. The approach described in this section applies to the continuation, cohort-tracking completion, progression and student experience indicators for which the outcome is observed at an individual student level before being aggregated to report on at provider level.

4. The confidence intervals which underpin the construction of the green shaded bars are created using the Jeffreys interval.[10] We have used the Jeffreys interval method because it has been shown to perform well in a wide range of circumstances in the assessment of many and diverse providers, including where the denominator is small, or the observed proportion is close to 0 per cent or 100 per cent.[11] The Jeffreys interval is calculated using the Jeffreys prior[12] for the binomial proportion, p, given n trials. Confidence intervals are calculated from the posterior distribution for p which is a Beta distribution with parameters $(np + 0.5, n - np + 0.5)$. In our case, p is the observed proportion and n is the denominator for the indicator in question. As the standard deviation of the binomial distribution decreases as the probability of success approaches 1 (i.e. an observed rate near 100 per cent), this results in a clear asymmetry in some of the bars.

5. To produce the figures in the supplementary table alongside the green shaded bar to inform our regulation of student outcomes we have determined the proportion of the distribution represented by the bar that falls above and below the numerical threshold. To do this, the cumulative distribution function (CDF) for the Jeffreys posterior distribution is used. The calculation is as follows:

[10] Jeffreys, Harold (1946). An invariant form for the prior probability in estimation problems. Proc. Royal Society, London. A186453–461. http://doi.org/10.1098/rspa.1946.0056.

[11] Brown et al (2001). Interval estimation for a binomial proportion Statistical Science. Vol. 16, No. 2, pages 101-133. http://dx.doi.org/10.1214/ss/1009213286.

[12] Although the Jeffreys interval has a Bayesian derivation it can also be justified from a frequentist perspective. See Brown et al (2001) – details in footnote 12.

a. **Proportion of the uncertainty distribution above the numerical threshold**: one minus the CDF at the numerical threshold.

b. **Proportion of the uncertainty distribution below the numerical threshold**: the CDF at the numerical threshold.

Approach for the compound completion indicator

6. The compound completion indicator is a cohort-based measure, rather than individual-based measure, because it relies on information from more than one population. It calculates six cohort withdrawal proportions which are added together and subtracted from 100 per cent to form the measure. Like the individual-based measures, in developing the approach to presenting uncertainty for this measure it is assumed that the probability of withdrawal for a student is equal to, and independent of, that of every other student in the same entry cohort. Under this assumption, the observed withdrawal proportion is binomially distributed. For other measures, we have taken the general approach to using the Jeffreys interval to calculate a confidence interval. However, because the compound completion indicator is a combination of multiple individual withdrawal proportions, each with its own degree of statistical uncertainty, we need to use a method that combines this uncertainty into an overall confidence interval.

7. As the Jeffreys interval method produces confidence intervals that cannot be feasibly combined, we instead approximate each binomial distribution using the normal distribution that has mean p and variance $p(1-p)/n$. Assuming that the probability of withdrawal for a student is independent of that of students in other cohorts, the variance can then be summed across each entry cohort that makes up the compound indicator calculation to estimate the variance of the indicator.

8. We have incorporated an adjustment to the normal approximation based on the Agresti-Coull interval[13], which has improved coverage probability compared with the normal approximation interval overall and in particular where the observed proportion is close to 0 per cent or 100 per cent. Like the Agresti-Coull interval, our approach adds pseudo-observations to the numerator and denominator of the observed proportion (for the confidence interval calculation only), the number of which varies with the chosen confidence level. As the compound completion indicator is a sum of proportions, we have split the pseudo-observations evenly across the proportions and adjusted the number used so that the total number is the same regardless of the number of summed proportions. For example, if only four of six cohort withdrawal proportions are available (because two of those cohorts don’t have any entrants), the number of pseudo-observations added are the same as when all six cohort withdrawal proportions are available.

9. The derivation of the confidence intervals used is below. In that derivation:

- $n_{ent,i}$ is the number of entrants in a given cohort, i

[13] Agresti and Coull (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. The American Statistician, Vol. 52, No. 2: pages 119–126; Agresti and Caffo (2000). Simple and Effective Confidence Intervals for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures. The American Statistician, Vol. 54, No. 4, pages 280-288.

- $n_{W,i}$ is the number of entrants that withdrew in the year relevant to the compound completion indicator in a given cohort, i

- Y is the number of cohorts with at least one entrant

- z is the quantile of the standard Normal corresponding to the desired confidence interval. For example, z ≈ 1.96 for the 95 per cent confidence interval.

$$\sum_{i=1}^{\mathrm{i=Y}} \frac{n_{W,i} + \frac{z^2}{2Y}}{n_{ent,i} + \frac{z^2}{Y}} \pm \mathrm{z} \times \sqrt{\sum_{\mathrm{i=1}}^{\mathrm{i=Y}} \frac{n_{W,i} + \frac{z^2}{2Y}}{\left(n_{ent,i} + \frac{z^2}{Y}\right)^2} \times \left(1 - \frac{n_{W,i} + \frac{z^2}{2Y}}{n_{ent,i} + \frac{z^2}{Y}}\right)}$$

10. To produce the figures in the supplementary table alongside the green shaded bar to inform our regulation of student outcomes we must adapt the approach used for the other indicators. This is because our method for deriving confidence intervals for the compound completion indicator is not based on a single underlying distribution but on multiple, slightly shifted normal distributions, one for each confidence interval. Our adapted approach is as follows. We first identify the confidence interval that corresponds to the numerical threshold (i.e. the one where the threshold is at either the lower or upper bound). Once we have identified this confidence interval, we determine the proportion of the associated normal distribution that is above and below the numerical threshold directly from the confidence level. For example:

    a. If the numerical threshold coincides with the lower limit of the 95 per cent confidence interval, then the proportion of the uncertainty distribution represented by the bar that is below the numerical threshold is taken to be 2.5 per cent and the proportion above the numerical threshold is taken to be 97.5 per cent.

    b. If the numerical threshold coincides with the upper limit of the 70 per cent confidence interval, then the proportion of the uncertainty distribution that is below the numerical threshold is taken to be 85 per cent and the proportion above the numerical threshold is taken to be 15 per cent.

11. In determining the statistical uncertainty for the compound completion indicator, in some extreme cases where the proportion is very large or very small or where the number of entrants in one or more academic years are small, it is possible for the confidence interval to be below 0 per cent or above 100 per cent. Where this occurs, they are marked and the shaded bars are truncated at 0 per cent and 100 per cent as appropriate. In these cases, the proportions of the uncertainty distribution above or below the numerical threshold should generally be considered as less reliable.

Annex B: Presenting uncertainty about relative performance

1. In presenting student outcomes and student experience indicators to inform our regulation of student outcomes and the TEF, a provider's relative performance is represented by blue shaded bars. They represent the statistical uncertainty associated with the observed value of the **difference** between a provider's indicator and its corresponding benchmark as a point estimate. This annex provides a full technical description of the statistical methods used to compute the confidence intervals which contribute to the construction of the blue shaded bars. It is aimed at readers with an in-depth knowledge of advanced statistical methods and assumes a familiarity with statistical formulae and notation.

2. The OfS uses benchmarking to create a comparator to absolute performance. The method to determine the benchmark and hence the difference between absolute performance and the benchmark follows the methodology described by Draper and Gittoes (2004)[14] and the most relevant elements of this methodology are described in Annex C of this document. The method includes a derivation of the standard deviation[15] of the difference between the absolute performance and the benchmark, which incorporates uncertainty in both components. They describe the relationship between the absolute performance and the benchmark and present evidence that the differences are normally distributed.

3. Each of the blue shaded bars represent a normal distribution with the distribution mean equal to the observed difference from benchmark and the distribution variance as the standard deviation squared. The distribution formula for the difference is:

$$N(\text{Difference}, (\text{Standard deviation})^2)$$

4. Where absolute performance is near 0 per cent or 100 per cent, it is possible for the distribution of the difference from benchmark represented by the blue shaded bar to imply that the absolute performance (i.e. if you centred this distribution around the observed absolute rate) could extend below 0 per cent or above 100 per cent. In constructing these bars, we have explicitly not adjusted for this and have instead tried to mitigate this issue by presenting the green shaded bar alongside it. This is because the green shaded bar, for all measures except the compound completion indicator, does not have this issue due to its derivation. The use of both charts reduces the risk that a user will misinterpret the uncertainty on the difference from benchmark in these cases.

5. To produce the figures in the supplementary table alongside the blue shaded bar we have determined the proportion of the distribution represented by the bar that falls around the numerical thresholds. To do this, the cumulative distribution function (CDF) for the normal distribution is used. To the left of the boundary of the numerical threshold the proportion is given by the CDF, to the right of the boundary of the numerical threshold the proportion is given

[14] Draper, D and Gittoes, M (2004). Statistical analysis of performance indicators in UK higher education. Journal of the Royal Statistical Society. Series A (Statistics in Society), 167, Part 3, pages 449-474.

[15] Because these are standard deviations of a statistic (the difference), they are more usually called standard errors.

by one minus the CDF. The numerical thresholds used differ between our regulation of student outcomes and the TEF:

a. **For our regulation of student outcomes**:

   i. **Proportion of the uncertainty distribution above the benchmark**: one minus the CDF at 0

   ii. **Proportion of the uncertainty distribution below the benchmark**: the CDF at 0.

b. **For the TEF**:

   i. **Proportion of the uncertainty distribution materially above benchmark**: one minus the CDF at 2.5

   ii. **Proportion of the uncertainty distribution materially below benchmark**: the CDF at -2.5

   iii. **Proportion of the uncertainty distribution broadly in line with benchmark**: one minus the sum of the results for materially above benchmark and materially below benchmark.

Annex C: Technical detail and formulae for calculating benchmarks

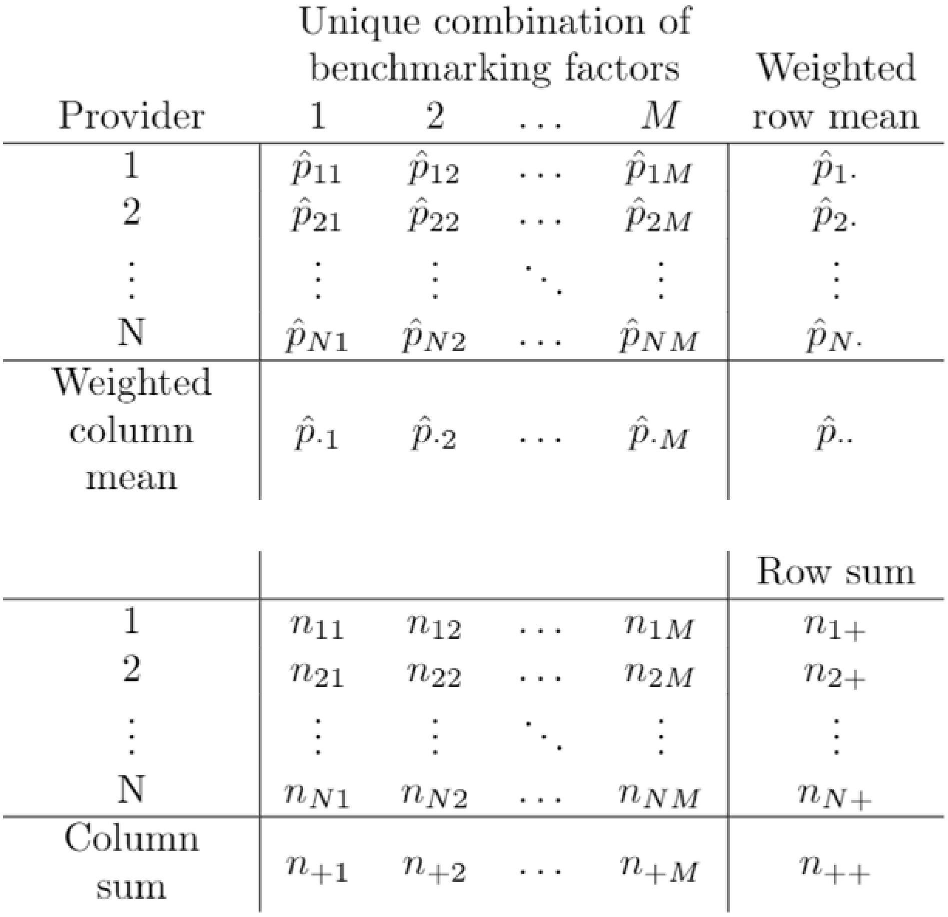
1. The general approach to benchmarking follows the design-based adjustment method described in 'Statistical analysis of performance indicators in UK higher education' by Draper and Gittoes (2004).[16] This annex summarises the key information from that methodology.

General approach

2. In this method, for each unique combination of benchmarking factors (described as potential confounding factors (PCFs) in the literature), an observed rate for the measure, and the number of students that inform it, is calculated for both the sector and each provider. The presentation of these rates and number of students for each unique combination of benchmarking factors can be visualised as two large grids as shown in Figure 1 below (the rates shown in the top table, with the number of students in the bottom table). In this figure, M represents the number of unique combinations of benchmarking factors. The method is based on a further cross-tabulation of the N providers by these M categories. The '.' and '+' notations in subscripts indicate averaging and summing over the relevant columns or rows of the table respectively. Within each table, each cell ij contains n_{ij} students from provider i with unique combination of benchmarking factors j. The observed rate of success of these students is $\hat{p}_{ij}$. Each weighted row mean, $\hat{p}_{i.}$ is the observed absolute performance for provider i and $\hat{p}_{.j}$ is the observed absolute performance for students with unique combination of benchmarking factor j across all students in the sector.

[16] Draper, D and Gittoes, M (2004). Statistical analysis of performance indicators in UK higher education. Journal of the Royal Statistical Society. Series A (Statistics in Society), 167, Part 3, pages 449-474.

Figure 1: A tabular presentation of the rates and number of students for each unique combination of benchmarking factors per provider



	Unique combination of benchmarking factors				Weighted
Provider	1	2	$\dots$	M	row mean
1	$\hat{p}_{11}$	$\hat{p}_{12}$	$\dots$	$\hat{p}_{1M}$	$\hat{p}_{1\cdot}$
2	$\hat{p}_{21}$	$\hat{p}_{22}$	$\dots$	$\hat{p}_{2M}$	$\hat{p}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
N	$\hat{p}_{N1}$	$\hat{p}_{N2}$	$\dots$	$\hat{p}_{NM}$	$\hat{p}_{N\cdot}$
Weighted column mean	$\hat{p}_{\cdot 1}$	$\hat{p}_{\cdot 2}$	$\dots$	$\hat{p}_{\cdot M}$	$\hat{p}_{\cdot\cdot}$

					Row sum
1	n_{11}	n_{12}	$\dots$	n_{1M}	n_{1+}
2	n_{21}	n_{22}	$\dots$	n_{2M}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
N	n_{N1}	n_{N2}	$\dots$	n_{NM}	n_{N+}
Column sum	n_{+1}	n_{+2}	$\dots$	n_{+M}	n_{++}

3. The observed absolute performance, $\hat{p}_{i.}$, for the provider can be directly read from the tables in Figure 1. The structure of the table allows us to consider the question: 'What would the observed absolute performance have been at provider i, if its distribution of students across the unique combination of benchmarking factors had been what is was, but its rates were replaced by the sector rates, $\hat{p}_{.j}$?'. These can be summarised as follows:

    a. The observed absolute performance, $\widehat{O}_i$, at provider i is:

    $$\hat{O}_i = \hat{p}_{i.} = \frac{1}{n_{i+}} \sum_{j=1}^{M} n_{ij}\hat{p}_{ij}$$

    b. The benchmark, $\hat{E}_i$, at provider i is:

    $$\hat{E}_i = \frac{1}{n_{i+}} \sum_{j=1}^{M} n_{ij}\hat{p}_{.j}$$

    c. The difference between the observed absolute performance and benchmark, $\widehat{D}_i$, at provider i is:

    $$\widehat{D}_i = \hat{O}_i - \hat{E}_i$$

4. To aid interpretation of the observed difference, the standard deviations of the differences between the absolute performance and benchmark have been calculated. A standard deviation measures the amount by which one would expect a statistic to change, based solely on random sampling. Because these are standard deviations of a statistic (the difference), they are more usually called standard errors.

5. To calculate the standard deviation, the formula for the difference is adjusted using algebraic manipulation (the full manipulation can be found in the literature) to be written as a weighted sum of all cells in the tables shown in Figure 1:

$$\widehat{D}_i = \sum_{j=1}^{M} \sum_{k=1}^{N} \lambda_{ikj} \hat{p}_{kj}$$

where $\lambda_{ikj} = \frac{n_{ij}}{n_{i+}} (\delta_{ik} - \frac{n_{kj}}{n_{+j}})$

and $\delta_{ik} = 1 \quad if\ i = k,$
$\delta_{ik} = 0 \quad if\ i \neq k$

Assuming the $\hat{p}_{kj}$ terms are independent, the variance is given by:

$$Var(\widehat{D}_i) = \sum_{j=1}^{M} \sum_{k=1}^{N} \lambda_{ikj}^2 Var(\hat{p}_{kj})$$

The literature shows that a reasonable estimate for the variance of $\hat{p}_{kj}$ can be made by using a shrinkage estimation procedure:

$$Var(\hat{p}_{kj}) = \frac{\hat{p}_{kj}^*(1 - \hat{p}_{kj}^*)}{n_{kj}}$$

where $\hat{p}_{kj}^* = 0.5\hat{p}_{..} + 0.5\hat{p}_{kj}$

and $\hat{p}_{..}$ is the overall rate of the sector.

The square root of the variance of $\widehat{D}_i$ gives the standard deviation.

6. We calculate the average contribution to benchmark for provider, i, using a similar weighted average calculation. This statistic calculates the contribution of the provider's own students on the sector averages that informs the calculation of the provider's benchmark of the form:

$$average\ contribution\ to\ the\ benchmark_i = \sum_{j=1}^{M} \frac{n_{ij}^2}{n_{+j} n_{i+}}$$

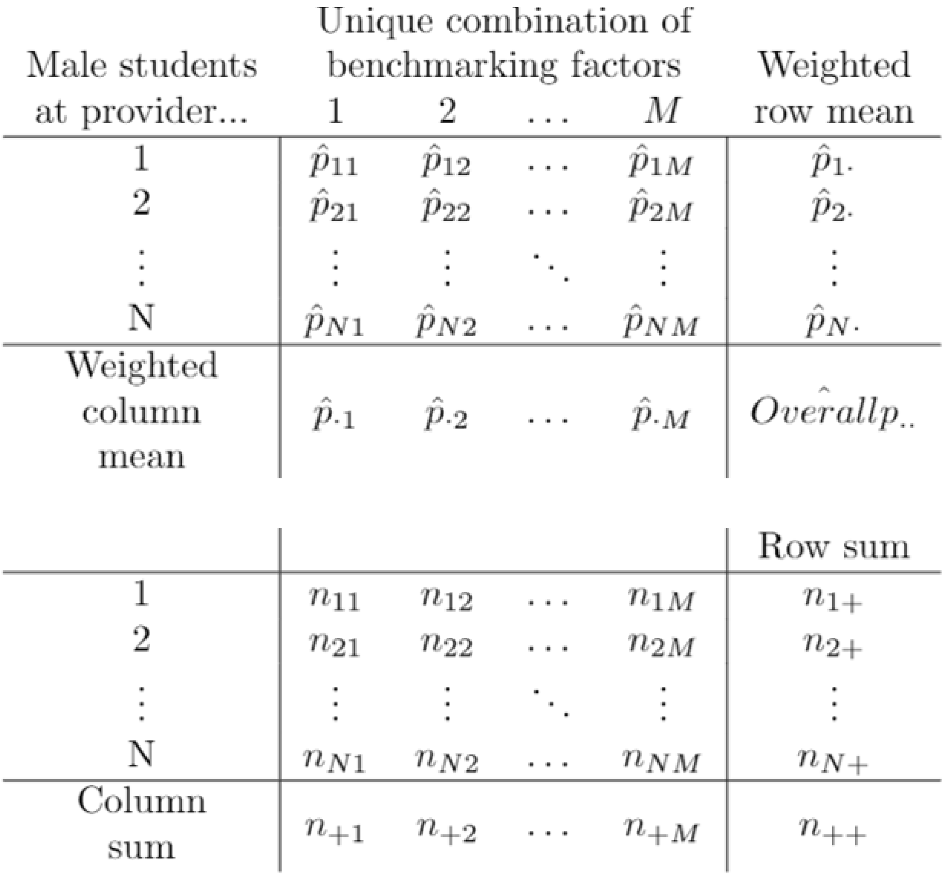
Benchmarking split indicators

7. In the calculation of the standard deviation for the purposes of benchmarking split indicators a small adjustment is made within the formulae described in the general approach above. The

approach to create an estimate for the variance of $\hat{p}_{kj}$ by using a shrinkage estimation is the same, but the value for $\hat{p}_{..}$ used in the derivation of $\hat{p}^*_{kj}$ remains the overall rate of the sector calculated at provider level. This is instead of using $\hat{p}_{..}$ created based on the subset of the provider and sector to the split indicator. This adjustment is made to ensure that the shrinkage estimation is applied consistently between the overall provider split indicator and other split indicators. For example, in a case where a provider delivers only a single subject, the standard deviation could appear different for the provider-level indicator and the split for the subject only because of the shrinkage estimation.

8. These differences in the approach to calculating benchmarks for split indicators is presented in the same tabular presentation as in Figure 1 in Figure 2, which assumes the split indicator being calculated is for ‘Male’ students. The $\hat{p}_{..}$ has been relabelled as $\widehat{Overallp}_{..}$. Otherwise, the notation is the same as described in paragraph 2.

Figure 2: A tabular presentation of the rates and number of students for each unique combination of benchmarking factors per provider for male students



	Unique combination of				
Male students	benchmarking factors				Weighted
at provider...	1	2	...	M	row mean
1	$\hat{p}_{11}$	$\hat{p}_{12}$	$\cdots$	$\hat{p}_{1M}$	$\hat{p}_{1\cdot}$
2	$\hat{p}_{21}$	$\hat{p}_{22}$	$\cdots$	$\hat{p}_{2M}$	$\hat{p}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
N	$\hat{p}_{N1}$	$\hat{p}_{N2}$	$\cdots$	$\hat{p}_{NM}$	$\hat{p}_{N\cdot}$
Weighted column mean	$\hat{p}_{\cdot 1}$	$\hat{p}_{\cdot 2}$	$\cdots$	$\hat{p}_{\cdot M}$	$\widehat{Overallp}_{..}$

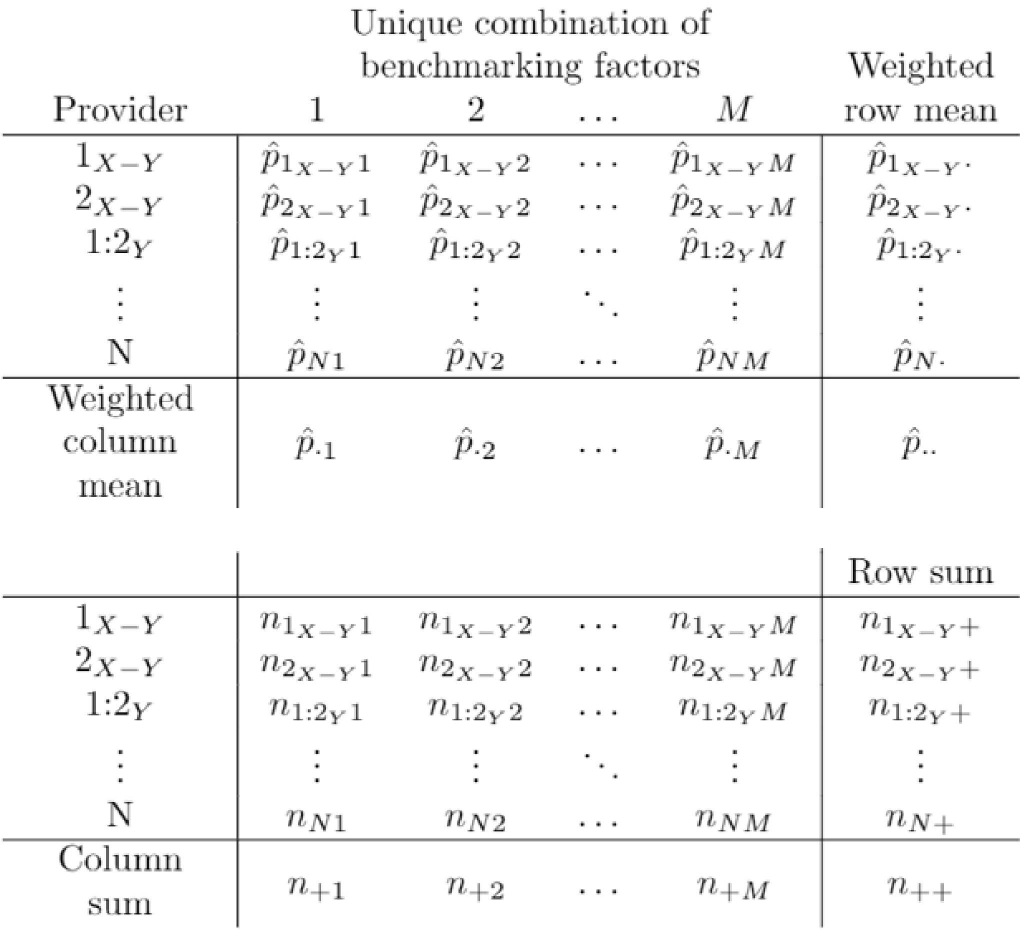
					Row sum
1	n_{11}	n_{12}	$\cdots$	n_{1M}	n_{1+}
2	n_{21}	n_{22}	$\cdots$	n_{2M}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
N	n_{N1}	n_{N2}	$\cdots$	n_{NM}	n_{N+}
Column sum	n_{+1}	n_{+2}	$\cdots$	n_{+M}	n_{++}

Adjustments to the general approach to benchmarking for the ‘taught or registered (TorR)’ population

9. We have proposed that in our regulation of student outcomes and the TEF, we will look at the population of students who are either registered or taught at the provider in question. This requires an adjustment to the approach used to construct benchmarks in this view. This is because students can be associated with more than one provider in the indicators. However, the benchmarking methodology assumes that students per provider per unique combination of benchmarking factors are independent from another combination.

10. The design-based adjustment methodology by Draper and Gittoes (2004) is adjusted as follows. We are placing any students that would be allowed to contribute to more than one provider in its own 'dummy' provider. These are students that contribute to the provider's indicator who registers them, but also to another provider's indicator who teaches them. To visualise this, the approach is presented in the same tabular presentation as in Figure 1 in Figure 3. In this figure, providers 1 and 2 share some duplicated students, *Y*, and their overall student population including these students is presented by *X*. The 'dummy provider' has been included as a separate row, shown as $1{:}2_Y$. Otherwise, the notation is the same as described in paragraph 2.

Figure 3: A tabular presentation of the rates and number of students for each unique combination of benchmarking factors per provider for the taught or registered population



	Unique combination of benchmarking factors				
Provider	1	2	. . .	M	Weighted row mean
1_{X-Y} 2_{X-Y} $1{:}2_Y$ ⋮ N	$\hat{p}_{1_{X-Y}1}$ $\hat{p}_{2_{X-Y}1}$ $\hat{p}_{1:2_Y1}$ ⋮ $\hat{p}_{N1}$	$\hat{p}_{1_{X-Y}2}$ $\hat{p}_{2_{X-Y}2}$ $\hat{p}_{1:2_Y2}$ ⋮ $\hat{p}_{N2}$	. . . . ⋱ . . .	$\hat{p}_{1_{X-Y}M}$ $\hat{p}_{2_{X-Y}M}$ $\hat{p}_{1:2_YM}$ ⋮ $\hat{p}_{NM}$	$\hat{p}_{1_{X-Y}\cdot}$ $\hat{p}_{2_{X-Y}\cdot}$ $\hat{p}_{1:2_Y\cdot}$ ⋮ $\hat{p}_{N\cdot}$
Weighted column mean	$\hat{p}_{\cdot 1}$	$\hat{p}_{\cdot 2}$	. . .	$\hat{p}_{\cdot M}$	$\hat{p}_{\cdot\cdot}$

					Row sum
1_{X-Y} 2_{X-Y} $1{:}2_Y$ ⋮ N	$n_{1_{X-Y}1}$ $n_{2_{X-Y}1}$ $n_{1:2_Y1}$ ⋮ n_{N1}	$n_{1_{X-Y}2}$ $n_{2_{X-Y}2}$ $n_{1:2_Y2}$ ⋮ n_{N2}	. . . . ⋱ . . .	$n_{1_{X-Y}M}$ $n_{2_{X-Y}M}$ $n_{1:2_YM}$ ⋮ n_{NM}	$n_{1_{X-Y}+}$ $n_{2_{X-Y}+}$ $n_{1:2_Y+}$ ⋮ n_{N+}
Column sum	n_{+1}	n_{+2}	. . .	n_{+M}	n_{++}

11. This manipulation to create a 'dummy' provider means:

a. The $\hat{p}_{kj}$ terms are independent across the whole grid because no students are duplicated within the grid.

b. There is no effect on the calculation of the sector average, $\hat{p}_{.M}$ because no students are duplicated within the grid.

c. The approach to estimating the variance of the difference used in the general approach for benchmarking can be used. In this example given in Figure 3, the variance of the difference is calculated for each provider, 1_{X-Y}, 2_{X-Y} and $1{:}2_Y$.

12. To calculate the difference and to estimate the variance per provider (including students that are duplicated across providers – in this example the variance for provider 1, rather than provider 1 without any students that are duplicated across providers), it is then necessary to combine the information calculated across the provider and any associated 'dummy' providers. Our derivation is as follows:

    Subscript Z represents the number of associated 'dummy' providers from provider i.

    n_{i_Z} represents the number of students from provider i, per 'dummy' provider Z.

    D_i represents the difference (indicator – benchmark) from provider i.

    D_{i_Z} represents the difference (indicator – benchmark) from provider i, per 'dummy' provider Z.

    The difference can be written as a weighted sum of the difference across multiple 'dummy' providers:

$$D_i = \frac{n_{i_1}D_{i_1} + n_{i_2}D_{i_2} + \cdots + n_{i_Z}D_{i_Z}}{n_{i_1} + n_{i_2} + \cdots + n_{i_Z}}$$

    Therefore, the variance of this weighted sum of difference is:

$$Var(D_i) = Var\left(\frac{n_{i_1}D_{i_1} + n_{i_2}D_{i_2} + \cdots + n_{i_Z}D_{i_Z}}{n_{i_1} + n_{i_2} + \cdots + n_{i_Z}}\right)$$

    This is equivalent to:

$$Var(D_i) = \left(\frac{1}{n_{i_1} + n_{i_2} + \cdots + n_{i_Z}}\right)^2 \times Var(n_{i_1}D_{i_1} + n_{i_2}D_{i_2} + \cdots + n_{i_Z}D_{i_Z})$$

$$Var(D_i) = \left(\frac{1}{n_{i_1} + n_{i_2} + \cdots + n_{i_Z}}\right)^2 \\ \times \left\{Var(n_{i_1}D_{i_1}) + Var(n_{i_2}D_{i_2}) + \cdots + Var(n_{i_Z}D_{i_Z}) + covariance\right\}$$

    As students do not appear more than once across Z 'dummy' providers, we can keep the assumption that the $\hat{p}_{kj}$ terms are independent. By combining 'dummy' providers we minimise the covariance between our differences, but inevitably there will a small amount of shared data[17], and hence covariance between them. In these calculations we are assuming that the covariance term is near zero. By also bringing out the n_i terms:

$$Var(D_i) = \left(\frac{1}{n_{i_1} + n_{i_2} + \cdots + n_{i_Z}}\right)^2 \times \left\{{n_{i_1}}^2 Var(D_{i_1}) + {n_{i_2}}^2 Var(D_{i_2}) + \cdots + {n_{i_Z}}^2 Var(D_{i_Z})\right\}$$

13. This derivation shows that we can estimate the variance for the entire provider by taking a weighted sum of the estimated variances for each of its 'dummy providers'. The square root of this variance gives the standard deviation.

[17] This only impacts the calculations of the uncertainty for the relative performance and not the absolute performance.

14. We have tested our assumption that the covariance term is near zero by comparing the standard deviations to the taught provider view (which does not need this adjustment because students are not duplicated across providers). We worked with the TEF metrics peer review group[18] to gain assurance over the low impact of some marginal differences in the standard deviations we identified.

15. We also adjust the calculation of the average contribution to benchmark for provider, i, using a similar weighted average calculation across dummy providers. This can be written as a weighted sum of the difference across multiple 'dummy' providers, where:

    C_i is the contribution to the benchmark from provider i.

    C_{i_Z} is the contribution to the benchmark from provider i, per 'dummy' provider Z.

    $$C_i = \frac{n_{i_1}C_{i_1} + n_{i_2}C_{i_2} + \cdots + n_{i_Z}C_{i_Z}}{n_{i_1} + n_{i_2} + \cdots + n_{i_Z}}$$

Benchmarking the compound completion indicator

16. The general approach to benchmarking assumes an individual-based measure is being calculated. However, the compound completion indicator is a cohort-based based measure rather than individual-based because it relies on information from more than one population. It calculates six cohort withdrawal proportions which are added together and subtracted from 100 per cent to form the measure.

17. To construct the benchmark for the compound completion indicator we consider each of the six entry cohorts as individual-based measures where we are measuring the withdrawal rate for each cohort. That means we can use the general approach to benchmarking to determine the withdrawal rates for students entering six years ago, separately to students entering five years ago, four years ago, three years ago and so on. This gives a benchmarked rate for each of the withdrawal periods. These benchmarked rates can be summed and subtracted from 100 per cent to give the benchmarked compound completion indicator, which mirrors the construction of the absolute performance for the indicator.

18. The benchmark can then be compared to the absolute performance for the indicator to create a difference. An estimate for the standard deviation of the difference would also be calculated following the general approach to benchmarking. To estimate the standard deviation, we calculate the square root of the sum of the variances of the difference for the six benchmarked withdrawal rates. This approach does not violate any underlying assumptions in the design-based adjustment method used by Draper and Gittoes (2004).

[18] Further details about the group are available at www.officeforstudents.org.uk/advice-and-guidance/teaching/future-of-the-tef/tef-metrics-peer-review-group/.

Annex D: Worked example of benchmarking calculations

19. This annex includes two fictional, simplified examples to demonstrate how we calculate benchmarks across our measures:

    a. Example 1 presents an example for calculating benchmarks for continuation measures. This example demonstrates the method that applies to the calculation of benchmarks for continuation, cohort-tracking completion, progression and student experience indicators for which the outcome is observed at an individual level before being aggregated to report on a provider.

    b. Example 2 presents an example for calculating benchmarks for the compound completion measure. Due to the construction of this measure, the methodology to calculating the benchmark is slightly different to that used for other measures.

Example 1

20. In this fictional, simplified example, assume that we are seeking to calculate benchmarks for continuation measures using only two benchmarking factors which affect the outcomes we are measuring. Specifically, we want to take account of students' age on entry to higher education, and the subject that they are studying. Suppose that students' age is defined as either 'young' or 'not young' and that the higher education sector delivers provision in only three subject areas (agriculture, maths and history).

21. That means that for this measure there are six possible distinct benchmarking groups, set out in the table below.

Step one: the provider

22. The provider for which we are calculating a benchmark has 1,090 students studying agriculture and maths. Table D1 shows the provider's students, split across the six benchmarking groups, and the continuation rate that we observe for each of these groups.

23. Overall, the provider has a continuation rate of 94.3 per cent. This is effectively a weighted average of the rates for each group.

24. Note that the provider's observed continuation rate for young maths students is particularly low (92.0 per cent) in comparison to the observed rate for other groups at the provider. This low continuation rate is outweighed by the larger number of students in groups with higher observed continuation rates, such as young agriculture students.

Table D1: Distribution of the provider's observed continuation rates across benchmarking groups

Age group	Subject group	Number of students	Students in the benchmarking group as a proportion of total students	Observed continuation rate
Young	Agriculture	500	45.9%	95.0%
Young	History	0	0.0%	N/A
Young	Maths	150	13.8%	92.0%
Not young	Agriculture	400	36.7%	94.0%
Not young	History	0	0.0%	N/A
Not young	Maths	40	3.7%	98.0%
				Provider indicator
Total		1,090	100%	94.3%

Step two: the sector

25. There are 210,500 full-time students across the whole sector, studying agriculture, maths and history. Table D2 shows the sector's students, split across the six benchmarking groups, and the continuation rate that we observe for each of these groups across the sector as a whole.

26. Overall, the sector has a continuation rate of 96.6 per cent.

27. Note that the sector's overall continuation rate is driven by high continuation rates observed for young history students (99.0 per cent), and the small student numbers for agriculture subjects, for which we observe relatively low rates for both young (95.0 per cent) and not young (94.0 per cent) students.

Table D2: Distribution of the sector's observed continuation rates across benchmarking groups

Age group	Subject group	Number of students	Observed continuation rate
Young	Agriculture	20,000	95.0%
Young	History	80,000	99.0%
Young	Maths	95,000	95.0%
Not young	Agriculture	5,000	94.0%
Not young	History	6,500	98.0%
Not young	Maths	4,000	98.0%
			Sector indicator
Total		210,500	96.6%

Step three: calculating the provider specific benchmark

28. So far, in Table D2, the sector's continuation rates are weighted against the numbers of students in the **sector** in each of the six distinct benchmarking groups. In Table D3 below, the sector's continuation rates are instead weighted to reflect the students in the **provider**.

29. Table D3 shows that weighting the sector's continuation rates by the proportion of students in each benchmarking group at the provider results in a weighted sector benchmark of 94.7 per cent for this provider.

30. This weighted sector rate is lower than the original sector rate shown in Table D2 since it no longer reflects the (relatively high) rates for history students (because the provider has no history students), and because the agriculture groups have a much higher weighting, reflecting that the provider has a higher proportion of agriculture students than the sector as a whole.

31. The provider's indicator (94.3 per cent) can now be compared with the weighted sector benchmark (94.7 per cent). The provider's rate is still lower than the rate observed for students with similar characteristics across the sector.

Table D3: Calculation of the provider benchmark using the sector's observed continuation rates across benchmarking groups

Age group	Subject group	Students in the benchmarking group as a proportion of total students at the provider (a)	Sector observed continuation rate (b)	Weighted sector continuation numbers (= a x b)
Young	Agriculture	45.9%	95.0%	43.6%
Young	History	0.0%	99.0%	0.0%
Young	Maths	13.8%	95.0%	13.1%
Not young	Agriculture	36.7%	94.0%	34.5%
Not young	History	0.0%	98.0%	0.0%
Not young	Maths	3.7%	98.0%	3.6%
Total		100%	Sector indicator 96.6%	**Provider benchmark** **94.7%** (= 43.6% + 0.0% + 13.1% + 34.5% + 0.0% + 3.6%)

Example 2

32. In this fictional, simplified example, assume that we are seeking to calculate benchmarks for the compound indicator. The compound completion indicator differs to other measures because it relies on information from more than one population. It calculates six cohort withdrawal proportions which are added together and subtracted from 100 per cent to form the measure. To construct the benchmark for the compound completion indicator we consider each of the six entry cohorts as individual-based measures where we are measuring the withdrawal rate for each cohort in the year in question. That means we determine a benchmark based on the withdrawal rates for students entering six years ago, separately to the benchmark for students entering five years ago, four years ago, three years ago and so on. This gives a benchmarked rate for each of the withdrawal periods. These benchmarked rates can be summed and subtracted from 100 per cent to give the benchmarked compound completion indicator, which mirrors the construction of the indicator.

33. In the example below, we are using only one benchmarking factor which affects the outcomes we are measuring. In this example we want to take account of students' age on entry to higher education which in this case is defined as either 'young' or 'not young'. That means that for this measure there are two possible distinct benchmarking groups, set out in the table below.

34. In essence the example and the approach to calculating the benchmark for the compound indicator is the same as other measures, but it is repeated six times for the different cohort withdrawal proportions. Throughout the example below, figures are shown up to one decimal place. When constructing the example, unrounded numbers were used.

Step one: the provider

35. In the given year the compound indicator is calculated, the provider for which we are calculating a benchmark had 470 students withdraw. The number of students that withdrew vary by entry cohort year.

36. Table D4 shows the provider's students, split across the two benchmarking groups for each of the six entry cohort years. For each entry cohort year, a 'Total' row is included which, in its final column, shows a weighted average of the rates for each group. At the bottom of the table, we demonstrate how the figures are used to construct the compound indicator by subtracting the withdrawal proportions per entry cohort year in the 'Total' rows from 100 per cent.

37. Overall, the provider has a compound indicator of 85.6 per cent.

Table D4: Distribution of the provider's withdrawal proportions across benchmarking groups per entry cohort year

Entry cohort year…	Benchmarking group	Students withdrawing per entry cohort year (x)	Total entrants per entry cohort year (y)	Students in the benchmarkin g group as a proportion of total entrants per entry cohort year	Cohort withdrawal proportion (= x / y)
One	Young	195	2,500	62.5%	7.8%
	Not young	80	1,500	37.5%	16%
	Total	**275**	**4,000**	**100%**	**6.9%**
Two	Young	70	2,000	80%	3.5%
	Not young	40	500	20%	8%
	Total	**110**	**2,500**	**100%**	**4.4%**
Three	Young	30	2,000	80%	1.5%
	Not young	0	500	20%	0%
	Total	**30**	**2,500**	**100%**	**1.2%**
Four	Young	25	2,000	80%	1.3%
	Not young	0	500	20%	0%
	Total	**25**	**2,500**	**100%**	**1%**
Five	Young	20	2,500	80%	0.8%
	Not young	0	625	20%	0%
	Total	**20**	**3,125**	**100%**	**0.6%**
Six	Young	10	2,500	80%	0.4%
	Not young	0	625	20%	0%
	Total	**10**	**3,125**	**100%**	**0.3%**
Provider compound indicator	**Total**	470	N/A	N/A	85.6% (=100 – 6.9 – 4.4 – 1.2 – 1 – 0.6 – 0.3)

Step two: the sector

38. Table D5 shows the sector's students, split across the benchmarking groups per entry cohort year, and the withdrawal proportions that we observe for each of these groups across the sector as a whole. It is constructed in the same way as Table D4 was for the provider, but illustrates data for the sector as a whole.

39. Overall, the sector has a compound indicator of 82.1 per cent.

Table D5: Distribution of the sector's withdrawal proportions across benchmarking groups per entry cohort year

Entry cohort year...	Benchmarking group	Students withdrawing per entry cohort year (x)	Total entrants per entry cohort year (y)	Cohort withdrawal proportion (= x / y)
One	Young	16,600	160,000	10.4%
	Not young	3,500	45,000	7.8%
	Total	**20,100**	**205,000**	**9.8%**
Two	Young	7,200	165,000	4.4%
	Not young	2,400	40,000	6.0%
	Total	**9,600**	**205,000**	**4.7%**
Three	Young	3,800	155,000	2.5%
	Not young	1,200	70,000	1.7%
	Total	**5,000**	**225,000**	**2.2%**
Four	Young	800	160,000	0.5%
	Not young	1,000	45,000	2.2%
	Total	**1,800**	**205,000**	**0.9%**
Five	Young	400	180,000	0.2%
	Not young	200	50,000	0.4%
	Total	**600**	**230,000**	**0.3%**
Six	Young	100	160,000	0.1%
	Not young	20	45,000	0.0%
	Total	**120**	**205,000**	**0.1%**
Sector compound indicator	**Total**	470	N/A	82.1% (=100 – 9.8 – 4.7 – 2.2 – 0.9 – 0.3 – 0.1)

Step three: calculating the provider specific benchmark

40. So far, in Table D5, the sector's withdrawal proportions are weighted against the numbers of students in the **sector** in each of the distinct benchmarking groups per entry cohort year. In Table D6 below, the sector's withdrawal proportions are instead weighted to reflect the students in the **provider** for the entry cohort year in question.

41. Table D6 shows that weighting the sector's withdrawal proportions by the proportion of students in each benchmarking group at the provider, for each entry cohort year in turn, gives a weighted sector benchmark rate for each of the entry cohort years. Summing these across the six entry cohort years and subtracting from 100 per cent results in a weighted sector benchmark of 82.4 per cent for the compound indicator for this provider.

42. The provider’s indicator (85.6 per cent) can now be compared with the weighted sector benchmark (82.4 per cent). The provider’s rate is higher than the rate observed for students with similar characteristics across the sector.

Table D6: Calculation of the provider benchmark using the sector’s observed withdrawal proportions across benchmarking groups per entry cohort year

Entry cohort year…	Benchmarking group	Students in the benchmarking group as a proportion of total entrants per entry cohort year at the provider (a)	Sector withdrawal proportion (b)	Weighted sector cohort withdrawal proportion (= a x b)
One	Young	62.5%	10.4%	6.5%
	Not young	37.5%	7.8%	2.9%
	Total	**N/A**	**N/A**	**9.4%**
Two	Young	80%	4.4%	3.5%
	Not young	20%	6.0%	1.2%
	Total	**N/A**	**4.7%**	**4.7%**
Three	Young	80%	2.5%	2.0%
	Not young	20%	1.7%	0.3%
	Total	**N/A**	**2.2%**	**2.3%**
Four	Young	80%	0.5%	0.4%
	Not young	20%	2.2%	0.4%
	Total	**N/A**	**0.9%**	**0.8%**
Five	Young	80%	0.2%	0.2%
	Not young	20%	0.4%	0.1%
	Total	**N/A**	**0.3%**	**0.3%**
Six	Young	80%	0.1%	0.1%
	Not young	20%	0.0%	0.0%
	Total	**N/A**	**0.1%**	**0.1%**
Provider benchmark compound indicator	**Total**	N/A	N/A	82.4% (=100 – 9.4 – 4.7 – 2.3 – 0.8 – 0.3 – 0.1)